DOCUMENT RESUME

ED 080 576 TM 003 092

AUTHOR	Houpt, Milton I.; Kress, Gerard
TITLE	Accuracy of Measurement of Clinical Performance in Dentistry.
SPONS AGENCY	Public Health Service (DHEW), Washington, D.C. Div. of Dental Health and Public Resources.
NOTE	37p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	College Students; College Teachers; Dentistry; Higher Education; *Performance Tests; *Rating Scales; *Student Evaluation; Test Reliability; Test Validity

ABSTRACT

This study was concerned with reliability and accuracy of measurement of clinical performance in operative dentistry. The influences on reliability and accuracy of the nature of the rating scale (that is, the number and the specificity of scale points), the extent of clinical experience of the rater, and the training of raters were investigated. Subjects included 30 instructors, 36 junior dental students, and 16 dental assistants from the University of Pittsburgh. In addition, five expert raters were used. The subject groups (instructors, students, assistants) were subdivided into three groups so that three different rating scales could be used. The scales were: a 2-point scale with two specified points, a 5-point scale with end points specified, and a 5-point scale with all points defined. At each of three sessions, subjects evaluated eight criteria of operative dentistry performance in five specimens, each containing one extracted mandibular second bicuspid. All scores were analyzed for between-judge and within-judge reliability, and for accuracy, that is, agreement with an expert score. Individual criterion scores as well as total test scores were used for the analysis. At one dental school instructors were able to reliably evaluate overall performance in operative dentistry. However, instructors were not able to assess specific criteria of performance accurately. Junior dental students performed similar to instructors, whereas dental assistants were able to be trained to rate overall performance fairly accurately. Practice with use of scale with immediate feedback did not result in high reliability and accuracy. (Author/DB)

ACCURACY OF MEASUREMENT OF CLINICAL PERFORMANCE IN DENTISTRY*

Milton I. Houpt, DDS, PhD**
Gerard Kress, PhD***

**Associate Professor and Chairman, Department of Pedodontics, New Jersey Dental School, Jersey City.

***Principal Research Associate, Director of Educational Research in Dentistry, Harvard School of Dental Medicine, Boston

*This project was performed at the University of Pittsburgh, School of Dental Medicine, and was supported in part by USPHS Grant No. 5T01-DHO-2001-04.
The authors wish to acknowledge the assistance of Drs. Thomas Zullo and Stanley Jacobs of the University of Pittsburgh, and Dr. Richard Mackenzie of the University of Florida in the design and implementation of this study.

I. INTRODUCTION

Formal education is predicated on evaluation, and the most common form of evaluation is the assessment of student learning. Students are evaluated in order to determine whether they have achieved some minimal proficiency. The student is tested and retested so that at the end of some period of time he may be promoted or failed. Concomitantly, evaluation may serve as a measure of the effectiveness of teaching. Teachers, teaching techniques and curricula are evaluated by reference to measures of student learning.

Evaluation presupposes valid measures of assessment. Nevertheless, there is little evidence of the validity of current methods in dentistry. When validity has been demonstrated, it has been based frequently on norm-referenced measures and has had little relation to external criteria. Such measures may well rank students one with another in regard to achievement, but they reveal little in regard to absolute level of individual achievement. In order to assess a student's performance judiciously, and to evaluate effects of teaching of clinical skills in dentistry, reliable and valid measures must be developed.

II. STATEMENT OF THE PROBLEM

This study was concerned with reliability and accuracy of measurement of clinical performance in operative dentistry. It investigated the influence on reliability and accuracy of the following independent variables: 1) the nature of the rating scale, that is, the number and the specificity of scale points; 2) the extent of clinical experience of the rater; and 3) the training of raters.

III. METHOD

A. Subjects

Subjects were selected from the School of Dental Medicine, University of Pittsburgh, sampling three populations, each with varying degrees of clinical experience: instructors, dental students and student dental assistants.

Thirty instructors were obtained from the Departments of Pedodontics and Restorative Dentistry including private practitioners and graduate students who were clinical instructors for at least six months. It was anticipated that instructors might have adopted individual preferences and biases in regard to operative dentistry because of their experience in practice and their training at different schools, and that such preferences might lead to lack of agreement in rating clinical performance. The instructors, therefore, were regarded as a somewhat heterogeneous group.

Thirty-six dental students were selected from the Junior class having had approximately six months of clinical experience in operative dentistry. The students in the top third of the class, as determined from the academic ranks of previous years, were selected as they were considered likely to have more knowledge of the subject than their classmates. Compared with the instructors, the students represented a relatively homogeneous group, having been trained in one environment.

The sixteen dental assistants chosen had one year or less of training, none of which included instruction in the cavity preparation or restoration procedures in operative dentistry.

The disparity in numbers of instructors, dental students and assistants was due to the numbers of individuals available for study.

Five full-time teachers were used for expert judgements. Of these, three are department chairmen at three different universities.

B. Instruments and Materials

Instruments. Three rating scales illustrating differences of numbers and specificity of scale points were used (Fig. I-III). Rating Scale A is a 2-point scale with end points partially specified. Rating Scale B is a 5-point scale with end points partially specified, and Rating Scale C is a 5-point scale with all points specified in detail. These scales were used to evaluate criteria of procedures most commonly performed by practicing dentists, i.e. cavity preparation, pulp protection and restoration of molar and bicuspid teeth.

For cavity preparation, the criteria "removal of fissures", "extension for prevention", "undermined enamel removed", "depth" and "retention" were selected. For pulp protection, the extent of coverage, and for restoration, the marginal integrity were chosen for assessment. These criteria are among those commonly used for assessment of operative dentistry performance as determined by Fernandez (1967) in a survey of dental schools in the United States and Canada. They also appear in textbooks of operative dentistry (Finn, 1967; Simon, 1956). Criterion definition for Scale C was arbitrarily determined from a consideration of the levels of error associated with each criterion.

Materials. The specimens which were evaluated by the subjects were plaster blocks (Fig. IV), each containing two extracted teeth, one mandibular first molar and one mandibular second bicuspid. The molar contained a class two cavity preparation (mesial-occlusal surface), and an amalgam restoration (buccal surface). The bicuspid contained a class two preparation in which the pulp protection (a white lining material) was placed. The specimens

RATING SCALE A Name ______________________________ Date ______________

Time Started ______________

Time Completed ______________

Mark "X" At The Appropriate Level Of Each Criterion For Each Tooth

		Tooth 1	Tooth 2	Tooth 3	Tooth 4	Tooth 5
Outline Form						
fissures removed	correct	1	1	1	1	1
	incorrect	0	0	0	0	0
extension for prevention	correct	1	1	1	1	1
(of proximal box walls)	incorrect	0	0	0	0	0
undermined enamel removed	correct	1	1	1	1	1
	incorrect	0	0	0	0	0
Depth	correct	1	1	1	1	1
pulpal floor	incorrect	0	0	0	0	0
axial wall	correct	1	1	1	1	1
	incorrect	0	0	0	0	0
Retention	correct	1	1	1	1	1
	incorrect	0	0	0	0	0
Pulp Protection	correct	1	1	1	1	1
	incorrect	0	0	0	0	0
Margins (of silver amalgam)	correct	1	1	1	1	1
	incorrect	0	0	0	0	0

Figure I. Rating Scale A, A Two Point Scale With Two Specified Points.

RATING SCALE B   Name ______________________ Date ________

Time Started ________

Time Completed ________

Mark "X" At The Appropriate Level Of Each Criterion For Each Tooth

		Tooth 1	Tooth 2	Tooth 3	Tooth 4	Tooth 5
Outline Form						
fissures removed	correct	4	4	4	4	4
		3	3	3	3	3
		2	2	2	2	2
		1	1	1	1	1
	incorrect	0	0	0	0	0
extension for prevention (of proximal box walls)	correct	4	4	4	4	4
		3	3	3	3	3
		2	2	2	2	2
		1	1	1	1	1
	incorrect	0	0	0	0	0
undermined enamel removed	correct	4	4	4	4	4
		3	3	3	3	3
		2	2	2	2	2
		1	1	1	1	1
	incorrect	0	0	0	0	0
Depth pulpal floor	correct	4	4	4	4	4
		3	3	3	3	3
		2	2	2	2	2
		1	1	1	1	1
	incorrect	0	0	0	0	0
axial wall	correct	4	4	4	4	4
		3	3	3	3	3
		2	2	2	2	2
		1	1	1	1	1
	incorrect	0	0	0	0	0
Retention	correct	4	4	4	4	4
		3	3	3	3	3
		2	2	2	2	2
		1	1	1	1	1
	incorrect	0	0	0	0	0
Pulp Protection	correct	4	4	4	4	4
		3	3	3	3	3
		2	2	2	2	2
		1	1	1	1	1
	incorrect	0	0	0	0	0
Margins (of silver amalgam)	correct	4	4	4	4	4
		3	3	3	3	3
		2	2	2	2	2
		1	1	1	1	1
	incorrect	0	0	0	0	0

Figure II. Rating Scale B, A Five Point Scale With Two Specified Points.

RATING SCALE C Name ______________________ Date __________

Time Started __________

Time Completed __________

Mark "X" At The Appropriate Level Of Each Criterion For Each Tooth

		Tooth 1	Tooth 2	Tooth 3	Tooth 4	Tooth 5
Outline Form						
fissures removed	-no fissures remaining	4	4	4	4	4
	-part of one fissure remaining	3	3	3	3	3
	-parts of more than one fissure remaining	2	2	2	2	2
	-one complete fissure remaining	1	1	1	1	1
	-more than one complete fissure remaining	0	0	0	0	0
extension for prevention (of proximal box walls)	-both walls extended	4	4	4	4	4
	-½ of one wall not extended	3	3	3	3	3
	-½ of both walls not extended	2	2	2	2	2
	-one complete wall not extended	1	1	1	1	1
	-both walls not extended	0	0	0	0	0
undermined enamel removed	-all undermined enamel removed	4	4	4	4	4
	-proximal wall undermined	3	3	3	3	3
	-occlusal ridge undermined	2	2	2	2	2
	-one cusp undermined	1	1	1	1	1
	-more than just a cusp undermined	0	0	0	0	0
Depth						
pulpal floor	-correct depth, 1mm into dentin	4	4	4	4	4
	-shallow, less than 1mm into dentin	3	3	3	3	3
	-too shallow, enamel remaining	2	2	2	2	2
	-very deep, more than 1mm into dentin	1	1	1	1	1
	-too deep, pulp exposure	0	0	0	0	0
axial wall	-correct depth, 1mm into dentin	4	4	4	4	4
	-shallow, less than 1mm into dentin	3	3	3	3	3
	-too shallow, enamel remaining	2	2	2	2	2
	-very deep, more than 1mm into dentin	1	1	1	1	1
	-too deep, pulp exposure	0	0	0	0	0
Retention	-retentive grooves in all required line angles	4	4	4	4	4
	-retentive grooves in proximal and ½ of occlusal	3	3	3	3	3
	-retentive grooves in ½ proximal and ½ occlusal	2	2	2	2	2
	-retentive grooves only in ½ of the proximal	1	1	1	1	1
	-no retentive grooves placed	0	0	0	0	0
Pulp Protection	-all exposed dentin covered	4	4	4	4	4
	-only pulpal floor and axial wall covered	3	3	3	3	3
	-½ of floor and all of axial wall covered	2	2	2	2	2
	-less than ½ of floor and all of axial wall covered	1	1	1	1	1
	-less than ½ of axial wall covered	0	0	0	0	0
Margins (of silver amalgam)	-all margins level with tooth surface	4	4	4	4	4
	-amalgam overextended	3	3	3	3	3
	-margin deficient at point in area of fissure	2	2	2	2	2
	-1mm or less of margin deficient	1	1	1	1	1
	-more than 1mm of margin deficient	0	0	0	0	0

Figure III. Rating Scale C, A Five Point Scale With Five Specified Points.

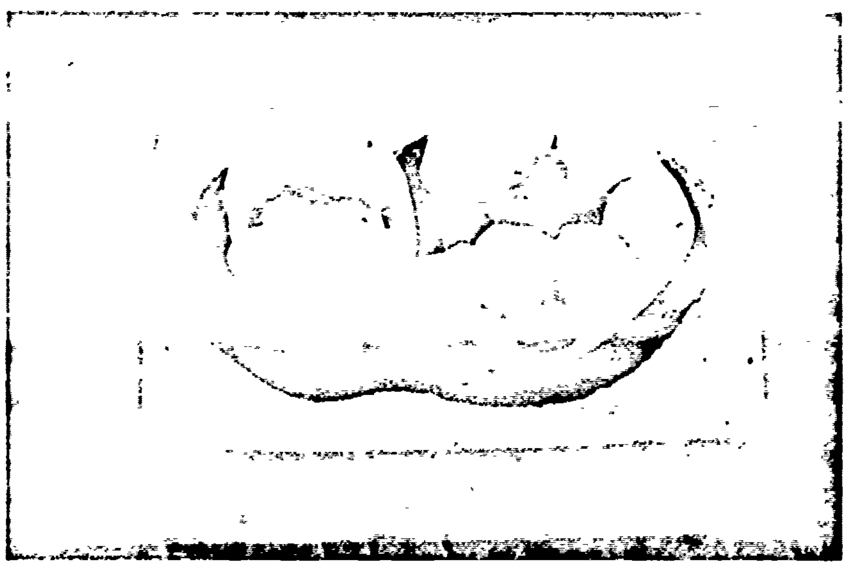

Figure IV. Mandibular First Molar And Second Bicuspid Teeth Mounted In A Plaster Block. Both Teeth Have Mesial-occlusal Cavity Preparations And The Molar Has A Buccal Amalgam.

were prepared so as to incorporate different levels of the criteria, randomly selected (using a table of random numbers) from all possible levels. Subjects participated during three sessions; therefore, three series of specimens were prepared so as to control for the effect of familiarity with the material. The three series of specimens, each containing five specimens, were prepared so as to be equivalent. The three series, therefore, were analagous to three parallel forms of a test. Each specimen was rated on eight criteria. Thus, when a series was rated, subjects assessed eight criteria on five specimens totalling forty judgements.

C. Experimental Design

The dependent variables of this study were reliability and accuracy of measurement. Reliability of measurement refers to both within-judge and between-judge reliability. Accuracy of measurement was operationally defined as degree of agreement with an expert score. Accuracy, thus, was a form of content validity.

The independent variables of the study were the nature of the rating scale, the extent of clinical experience of the rater, and the training of the raters.

In order to study the effects of the independent variables, each of the three samples of subjects (instructors, dental students, and assistants), with differing degrees of clinical experience were subdivided randomly into three groups (Groups A, B and C) yielding a total of nine groups (Instructors' Groups A, B and C, Students' Groups A, B and C, and Assistants' Groups A, B and C). The three different rating scales were used by the different groups, Scale A by Groups A, Scale B by Groups B, and Scale C by Groups C. All groups performed a rating task on three occasions (Observations 1, 2 and 3) separated by approximately 2-week intervals.

Observation 1 was made prior to any training, whereas Observations 2 and 3 each followed a training session. Thus a 3x3x3 factorial design resulted. The rating task involved the assessment of performance in operative dentistry, i.e. cavity preparation, pulp protection and restoration.

D. Procedure

1. The procedures which are performed most commonly by practicing dentists were chosen for the rating task i.e. cavity preparation, pulp protection and restoration.

2. Three different rating scales were developed incorporating criteria derived from a review of criteria used in dental schools of the United States and Canada. The criteria were arbitrarily defined for the 5-point scale from a consideration of the levels of error associated with each criterion.

3. Three series of five specimens were prepared incorporating levels of the criteria randomly chosen from a list of all possible levels of the criteria.

4. An expert score was derived from the majority or mode judgements of five full-time teachers for comparison with subjects' ratings. The experts had access to a magnifying glass and a millimeter rule when rating the teeth, although these were used only occasionally. The experts used Rating Scale C. When required for comparison with subjects' ratings with Scale A, a 2-point scale, expert judgements from the 5-point scale (Scale C) were dichotomized. The expert scores were dichotomized between points 3 and 4 of the 0, 1, 2, 3, 4 scale, as points 0, 1, 2 and 3 represented degrees of error whereas point 4 represented correctness.

5. Subjects viewed the teeth with the use of a standard dental operating light, and assessed each specimen independently of the others. No dental instruments were allowed. There was no time limit but the complete rating task usually lasted approximately ten minutes. Some technical terms were illustrated for the dental assistants at the beginning of the study but criteria were not specifically defined.

6. The subjects rated a separate series of five specimens on each of three occasions (Observations 1, 2 and 3) using one of the three rating scales. During Observations 2 and 3 training was administered.

7. The training consisted of performance with immediate feedback. Subjects rated specimens and after each rating they were provided with the correct rating as defined by the consensus score of expert raters. For those subjects using a 5-point scale (Scale B or C), the consensus score was one of five points, however, for those using a 2-point scale (Scale A) the consensus score was dichotomized so as to be one of two points. During the first session, the first series of five specimens was rated and no training was given. At the second session the first series was regraded. Two specimens of the first series were graded and correct ratings were provided. The subjects were told that the correct ratings were consensus ratings of all instructors rather than of experts only, as it was anticipated that individuals would more readily accept group opinion. As each criterion of the remaining three specimens of the first series was graded, correct ratings were provided. This procedure was facilitated by small booklets containing the correct ratings with one rating per page. After each judgement was made, the page

was turned indicating the correct judgement. The various groups of subjects used appropriate booklets with ratings corresponding to Scales A, B or C. After the first series was graded during the second session, the second ser [illegible]cimens was evaluated. At the third session, two specimens of the second series were regraded. Correct ratings were then provided, in addition to descriptions of the various factors which determined the particular ratings. The third series of specimens was then graded.

IV. RESULTS

The results of this study may be examined in two different ways. Either the total scores for all criteria or the individual criterion scores given by the raters to each specimen may be analyzed. The total score for all criteria is analagous to the total test score for a test with many items. The individual criterion score is analagous to a score for a particular item, one of many items in a total test. Total scores are used when items are relatively homogeneous. Individual item scores are considered when items are heterogeneous and measure different attributes. For this study, the data was analyzed using both the scores for individual criteria, and the scores for all criteria representing total test scores.

A. Within-Judge Reliability

Within-judge reliability was estimated from the repeated ratings of four specimens, two before training (rated during Observation 1 and 2), and two after training (rated during Observation 2 and 3). Judgements

made at Observation 1 were compared with judgements made at Observation 2 (before training) and judgements made at Observation 2 were correlated with judgements made at Observation 3 (after training). For each group, the 64 judgements (eight judgements each for two specimens repeated twice before training and twice after training) of each rater were analyzed to derive group reliability estimates before and after training. For Groups A the percent agreement statistic was used for individual criterion scores, whereas for Groups B and C the Pearson correlation coefficient was used. The percent agreement statistic was used because the Pearson coefficient and its derivatives produced spurious results for scores with a range of 2 when agreement was perfect or near perfect. Moderate to high reliability was arbitrarily defined as .70 for the Pearson correlation coefficient and 80 for the percent agreement. The Chi-Square test with the Yates' continuity correction for small values was used to test for significant differences between reliability estimates. In all instances a 2x2 comparison was made and the degree of freedom was one.

Within-judge reliability for all subjects before and after the first training session is reported in Table 1. For scores of all criteria (total test score), ten of 18 reliability estimates were moderate to high. For individual criterion scores, 59 of the 144 reliability estimates were moderate to high (percent agreement 80 or greater, Pearson correlation coefficients .70 or greater).

Subjects. For total test scores, the correlation coefficients for instructors ranged from .57 to .80 before training and from .45 to .73

TABLE 1

WITHIN-JUDGE RELIABILITY* BEFORE AND AFTER TRAINING

Treatment Group	Individual Criterion Scores 1	2	3	4	5	6	7	8	Total Scores
Before Training									
Instructors									
Scale A	80**	85	95	70	90	65	70	80	.57
Scale B	.35	.62	.69	.74	.71	.70	.62	.25	.77
Scale C	.37	.63	.70	.36	.04	.46	.82	.40	.80
Students									
Scale A	83	83	50	66	75	83	33	42	.75
Scale B	.48	.58	.61	.33	.72	.51	.74	.49	.76
Scale C	.57	.45	.61	.27	.53	.72	.85	.41	.53
Assistants									
Scale A	83	75	58	75	50	66	66	58	.80
Scale B	.66	.54	.0	.0	.73	.44	.28	.25	.63
Scale C	.50	.64	.42	.0	.0	.60	.87	.0	.36
After Training									
Instructors									
Scale A	100	95	90	85	90	80	85	85	.72
Scale B	.72	.46	.42	.45	.78	.60	.88	.74	.73
Scale C	.54	.74	.29	.05	.0	.52	.92	.66	.45
Students									
Scale A	91	100	91	58	91	100	42	66	.62
Scale B	.37	.76	.36	.39	.52	.47	.82	.78	.57
Scale C	.65	.07	.49	.62	.0	.52	.94	.73	.72
Assistants									
Scale A	91	91	83	83	50	91	83	83	.34
Scale B	.74	.73	.29	.01	.92	.78	.52	.83	.79
Scale C	.66	.57	.80	.21	.75	.42	.67	.83	.75

*Figures for Scale A for individual criterion scores are percents agreement. All other figures are Pearson correlation coefficients.

**Underlined values represent moderate to high reliability estimates (80 or greater for percents agreement, and .70 or greater for Pearson coefficients).

after training. The coefficients were high (.70 or greater) when instructors used Scales B and C before training and Scales A and B after training. For students the coefficients ranged from .53 to .76 and were high with Scales A and B before training. After training the coefficients ranged from .57 to .72 and were high only with Scale C. Assistants had high coefficients with Scale A (coefficients ranged from .36 to .80) before training, and with Scales B and C (coefficients ranged from .34 to .79) after training.

For individual criterion scores, the correlation coefficients for instructors ranged from .04 to .82 before training and from .0 to .92 after training. The percent agreement ranged from 65 to 95 before training and 80 to 100 after training. Before training, good reliability (.70 or greater) was evident in only 5 of 16 coefficients. After training, six coefficients were .70 or greater, although in only two cases were these the same scales for the same criteria which produced high coefficients before training. Five of 8 percents agreement were 80 or greater before training and after training all 8 percents agreement were 80 or greater. The number of high values after training (14 of 24) was not significantly different to the number before training (10 to 24) ($\chi^2 = .75$, $p > .05$).

The correlation coefficients for individual criterion scores of students ranged from .27 to .85 before training and from .0 to .94 after training. The percent agreement ranged from 42 to 83 before training and 42 to 100 after training. Before training, only 4 of 16 coefficients were .70 or greater and after training 5 were .70 or greater. In only two cases were the coefficients for specific criteria rated with a specific scale .70 or higher both before and after training. Before training 3 of 8

percents agreement were 80 or greater. These remained high after training, when five percents agreement were 80 or greater. The number of high values after training (10 of 24) was not significantly different to the number before training (7 of 24), ($\chi^2 = .36$, $p > .05$). Similarly there were no significant differences between numbers of high values for students and those for instructors ($\chi^2 = .36$ and $.75$, $p > .05$).

Reliability estimates for individual criterion scores for assistants were low before training as the correlation coefficient ranged from .0 to .87 with only 2 of 16 having been .70 or greater, and the percents agreement ranged from 50 to 83 and only 1 was 80 or greater. After training, however, the number of high reliability indices were significantly greater ($\chi^2 = 10.75$, $p < .01$). Eight of 16 coefficients were .70 or higher (range .01 to .92) and 7 of 8 percents agreement were 80 or greater (range 50 to 91).

Scales. For total test scores, subjects produced high indices in 10 of 18 instances with little difference between scales (three with Scale A, four with Scale B, and three with Scale C). For individual criterion scores, subjects with Scale A produced high indices in 29 of 48 instances before and after training. These subjects were significantly better than Scale C subjects who produced high indices in only 12 of 48 instances ($\chi^2 = 8.26$, $p < .01$). There was no significant difference in the performance of subjects with Scale B (18 high indices) as compared with Scale A subjects ($\chi^2 = 3.48$, $p > .05$) and Scale C subjects ($\chi^2 = .22$, $p > .05$).

Training. For total test scores both before and after training 5 of 9 indices were high. Four of the 9 indices changed from low to high (.70 or greater), 4 indices changed from high to low and 1 remained the same. With individual criterion scores before training only 20 of 72

indices were high. (The 72 indices were derived for each of 8 criteria rated with each of 3 scales, by 3 samples of subjects). After training, 39 of 72 indices were high indicating that, in general, training had the effect of raising internal consistency ($\chi^2 = 9.30$, $p < .01$). Three-quarters of the indices which were high before training remained high after training indicating that training generally did not destroy internal consistency in the rating of individual criteria. However, the marked effect of training on within-judge reliability was present only with the assistants and not with the instructors or students. Scale A subjects demonstrated greatest improvement in within-judge reliability for individual criterion scores with training, producing high indices in 9 of 24 instances before training and 20 of 24 instances after training ($\chi^2 = 8.71$, $p < .01$). Scale B and C subjects with 6 and 5 high indices of a maximum of 24 before training and 12 and 9 indices after training, demonstrated no significant differences with training ($\chi^2 = 2.2$ and $.90$, $p > .05$).

<u>Criteria</u>. Some criteria were rated with much lower within-judge reliability than others. "Undermined enamel removed" and "pulpal floor depth" were rated inconsistently both before and after training. The ratings of some criteria benefited more than others from training. "Extension for prevention", "pulp protection" and "margins" were rated much more consistently after training than before.

B. Between-Judge Reliability

Between-judge reliability was estimated for each cell in the 3x3x3 factorial design, that is, for each group of each population at each observation. It was calculated using both the total scores and the individual criterion scores given by raters. Finn's (1970) derivation of Ebel's (1951) analysis of variance techniques was used to derive indices of between-judge reliability using raw scores from the rating scales. Between-judge reliability estimates for all groups of subjects are reported in Table 2. When total scores were used, relatively good between-judge reliability was demonstrated (22 of 27 indices were high). However, when individual criterion scores were used, only 28 of 216 estimates were high (.70 or greater) indicating generally low agreement between judges when rating specific criteria.

Subjects. For total scores, reliability estimates ranged from .62 to .93 for instructors, .56 to .94 for students and .45 to .90 for assistants. Instructors and students had 8 of 9 indices high, whereas assistants had only 6 high indices, indicating little difference between instructors and students, but less reliability with assistants. When individual criterion scores were used, reliability estimates for instructors ranged from .0 to .93 with only 12 of 72 equal to or greater than .70. Estimates for students ranged from .09 to .96 with 14 equal to or greater than .70, and for assistants the range was .0 to .73 with only 2 estimates equal to or greater than .70. There was no significant difference in the performance of instructors and students ($\chi^2 = .52$, $p > .05$), but students and instructors performed significantly better than assistants ($\chi^2 = 4.11$, $.05 > p > .01$).

TABLE 2

BETWEEN-JUDGE RELIABILITY (FINN TECHNIQUE)

Treatment Group	1	2	3	4	5	6	7	8	Total Scores
Session One									
Instructors									
Scale A	.73*	.35	.49	.19	.45	.21	.08	.40	.62
Scale B	.62	.33	.38	.20	.25	.48	.03	.34	.75
Scale C	.48	.0	.47	.40	.84	.0	.72	.32	.87
Students									
Scale A	.74	.58	.12	.36	.22	.09	.12	.34	.56
Scale B	.63	.56	.55	.63	.57	.58	.48	.62	.81
Scale C	.37	.48	.33	.42	.63	.11	.78	.50	.88
Assistants									
Scale A	.28	.01	.0	.31	.0	.28	.09	.20	.45
Scale B	.44	.60	.43	.61	.33	.49	.17	.18	.61
Scale C	.18	.0	.0	.07	.0	.0	.05	.44	.75
Session Two									
Instructors									
Scale A	.63	.42	.67	.38	.76	.42	.31	.44	.75
Scale B	.43	.56	.33	.64	.39	.46	.07	.58	.80
Scale C	.56	.51	.83	.54	.80	.22	.90	.70	.93
Students									
Scale A	.58	.80	.48	.41	.58	.32	.38	.61	.75
Scale B	.65	.76	.66	.61	.58	.36	.59	.80	.90
Scale C	.26	.57	.71	.67	.86	.34	.96	.58	.94
Assistants									
Scale A	.44	.15	.55	.28	.41	.41	.39	.52	.75
Scale B	.45	.57	.56	.50	.70	.41	.62	.55	.90
Scale C	.01	.02	.0	.44	.0	.0	.59	.29	.75
Session Three									
Instructors									
Scale A	.40	.84	.56	.62	.59	.41	.39	.73	.85
Scale B	.46	.35	.22	.46	.46	.47	.49	.32	.73
Scale C	.67	.34	.44	.43	.70	.19	.93	.55	.91
Students									
Scale A	.33	.81	.36	.38	.55	.40	.42	.45	.73
Scale B	.54	.73	.62	.64	.42	.49	.77	.58	.81
Scale C	.41	.26	.88	.49	.89	.40	.95	.31	.91
Assistants									
Scale A	.44	.31	.12	.36	.36	.47	.44	.73	.55
Scale B	.50	.52	.22	.33	.44	.55	.65	.58	.76
Scale C	.43	.03	.0	.37	.0	.67	.63	.13	.71

*Underlined values represent moderate to high reliability estimates (.70 or greater).

Scales. For total scores almost all reliability estimates from Scales B and C were high (8 of 9 for Scale B and all 9 for Scale C) whereas for Scale A only one-half of the estimates were high (5 of 9). For individual criteria scores, there were no significant differences in numbers of high between-judge reliability indices produced by Scale A, B and C subjects. Scale B subjects had only 4 of 72 indices equal to or greater than .70 and Scale A and C subjects produced 8 and 15 moderate-to-high indices, respectively.

Training. With estimates for total scores, training increased reliability (5 of 9 high estimates, to 9 of 9, to 8 of 9). However, with individual criterion scores, training had little effect on between-judge reliability. Whereas before training 5 of 72 were .70 or greater, after the first training session only 12 indices were .70 or greater and after the second training session only 11 indices were of that magnitude. No significant differences were evident after the first training session ($\chi^2 = 1.35$, $p > .05$) or after the second training session ($\chi^2 = .0$, $p > .05$).

Criteria. Most criteria were unreliably rated (19 to 21 indices of 21 were low), however, 3 had indices greater than the others (5 to 7 of 21 were equal to or greater than .70). The reliably rated criteria were "extension for prevention", "axial wall depth", and "pulp protection".

C. Accuracy

In this study, accuracy was defined as degree of agreement with an expert score. In order to determine the effects on accuracy of the independent variables (subject, rating scale, and training) analyses of

variance were performed using absolute deviation scores (deviation of subject score from expert score). Table 3 lists the mean deviation scores for all subjects rating all scales in rating five models. Mean deviation scores for individual criteria were derived by comparison of rater individual criterion scores with those of the experts, whereas total scores were derived by comparison of the total scores assigned to specimens by raters and experts. The sum column of Table 3 represents the sum of the mean deviation scores for all of the eight individual criteria. Comparison of the sum scores with the total scores demonstrates a much greater degree of agreement between raters and experts (small deviation scores) when total specimen scores are considered rather than individual criterion scores.

An unweighted means analysis was used for the analyses of variance as cell frequencies were unequal. The F-ratios derived in the analyses are reported in Table 4.

<u>Subjects</u>. Inspection of Tables 3 and 4 reveals consistent effects of degree of experience on accuracy, that is, instructors and students were similar in regard to accuracy whereas assistants, whose deviation scores are about one-third greater than those of the instructors and students, were least accurate.

<u>Scales</u>. Accuracy in the use of a particular type of rating scale can be determined from examination of Tables 3 and 4. Table 4 demonstrates significant differences with all criteria. In all cases, Scale . [illegible] s to be most accurate and Scale C appears to be more accurate than Scale B (Table 3).

TABLE 3

MEAN DEVIATION SCORES* FOR ALL SUBJECTS USING ALL SCALES RATING FIVE MODELS

Group	Observation	Scale	Individual Criterion Scores 1	2	3	4	5	6	7	8	Sum**	Total*** Scores
Students	1	A	.4	.7	1.5	1.3	1.5	2.6	2.5	1.3	11.8	1.8
	2	A	.7	.3	.7	1.3	.5	2.0	1.8	.7	8.0	1.3
	3	A	1.2	.3	.9	1.3	.7	1.0	1.3	1.0	7.7	1.0
	1	B	7.7	4.1	4.7	6.1	5.3	6.3	5.7	6.0	45.9	3.5
	2	B	5.3	4.1	4.4	5.4	6.2	7.3	4.5	1.8	39.0	2.7
	3	B	5.3	3.5	4.7	7.1	6.5	6.3	4.1	4.8	42.3	3.7
	1	C	7.3	4.8	4.6	3.5	3.1	5.9	1.8	4.1	35.1	3.0
	2	C	7.0	5.4	1.3	3.2	1.8	5.1	.6	3.0	27.4	2.1
	3	C	6.8	4.9	1.2	3.0	.8	4.8	1.3	4.4	27.2	3.4
Instructors	1	A	.8	1.6	1.3	1.3	.8	2.5	2.4	1.3	12.0	1.9
	2	A	.5	.8	.4	1.1	.3	1.4	2.0	.9	7.4	1.2
	3	A	1.3	.2	.7	.7	.7	1.2	1.3	.2	6.3	0.7
	1	B	8.9	4.2	6.7	6.8	5.5	5.4	6.2	6.0	49.7	4.1
	2	B	4.9	4.2	5.3	6.7	6.9	5.9	6.0	3.7	43.6	4.4
	3	B	6.8	5.0	5.6	5.9	7.8	5.8	4.3	5.4	46.6	4.9
	1	C	8.3	5.2	4.1	5.4	1.5	6.8	2.6	4.4	38.3	2.8
	2	C	5.0	4.4	1.2	2.5	2.0	5.4	1.1	2.9	24.5	2.3
	3	C	5.5	4.1	2.5	4.9	1.7	5.7	1.5	3.9	29.8	3.5
Assistants	1	A	2.2	3.2	2.3	1.3	2.0	1.8	2.2	2.3	17.3	2.8
	2	A	1.2	1.7	.8	2.2	1.0	1.0	.8	.7	9.4	1.3
	3	A	.8	1.3	1.5	1.7	1.3	1.3	1.5	.3	9.7	1.1
	1	B	9.8	5.0	8.0	7.2	7.4	6.2	5.4	9.2	58.2	5.7
	2	B	6.6	9.4	3.8	8.0	5.8	7.4	5.2	4.0	50.2	2.9
	3	B	8.0	8.8	4.8	8.4	5.0	7.8	4.8	3.0	50.6	5.0
	1	C	10.6	6.4	8.4	6.6	4.6	8.2	4.0	8.4	57.2	4.9
	2	C	6.0	6.2	7.2	4.6	7.2	9.0	2.8	3.2	46.2	3.6
	3	C	5.8	9.3	6.8	2.8	5.8	9.5	2.5	3.8	46.3	4.5
Criterion Totals			134.7	109.1	95.4	110.3	93.7	133.6	80.2	90.7	847.7	

*For individual criteria, groups using Scale A could have a maximum score of 5, and those using Scales B and C, a maximum score of 20.

**This column represents the sum of the mean deviation scores for all of the eight individual criteria.

***This column represents the mean deviation scores when the total specimen scores of raters and experts were considered.

TABLE 4

F-RATIOS* FOR CRITERIA 1 TO 8 USING ALL SCALES AND ALL SUBJECTS

	df	1	2	3	4	5	6	7	8	Total
Subjects	2	5.70**	31.64	27.77	7.00	22.44	11.61	2.52	(4.46)***	87.16
Rating Scale	2	199.84	98.83	96.78	144.18	216.56	168.94	87.19	94.15	879.99
Observation	2	19.93	.23	19.15	1.38	.29	.35	8.73	34.40	41.08
Subjects X Scale	4	1.40	2.39	16.55	1.18	18.37	9.12	(3.33)	1.26	18.85
Subjects X Observation	4	2.41	1.91	.49	.96	1.98	1.75	.31	8.83	.96
Scale X Observation	4	4.25	(2.99)	1.27	(2.60)	2.11	2.00	.72	4.16	1.96
Subjects X Scale X Observation	8	.95	(2.12)	1.73	(2.00)	3.80	.22	.47	1.47	.41
Within Replicates	218									

* The F-ratio was calculated by using the within replicates mean square for the denominator.

** Underlined values are significant at the 1% level of probability.

*** Bracketed values are significant at the 5% level of probability.

However, with Scale A (the 2-point scale) the maximum deviation score possible is five, whereas with Scales B and C (the 5-point scales) the maximum deviation score is 20. Scale A, therefore, cannot be directly compared with Scales B and C with regard to accuracy of measurement. There is, however, a method of indirectly equating the scales for comparison. Scale A, the 2-point scale, could deviate by only one point, that is, the maximum deviation and the average maximum deviation from the true score would be one. Whenever the 2-point scale is used, there is either no deviation from the true score, or there is a deviation of one point. With the 5-point scale, the maximum deviation would be four points, for example, when the rating is 0 and the true score is 4. However, if the true score happened to be 2, then the maximum deviation could only be 2. Assuming that the expert scores represent the true scores, the expert scores for all criteria can be examined to determine whether the maximum deviation would be 2, 3 or 4. The average maximum deviation of rater scores was calculated to be 3.3 for the 5-point scales. Therefore, a rater using Scales A and B or C, and operating at the same level of performance with both scales, would be expected to deviate 3.3 times greater with the 5-point scale than with the 2-point scale. If all scores obtained with Scale A were to be multiplied by 3.3, comparison of the resulting values could be made with those of Scales B and C. Examination of the sum column of Table 3 demonstrates that when scores for Scale A are multiplied by 3.3 and compared with scores of Scales B and C, Scale A is more accurate (less deviant) than Scales B and C.

Training. The effect of training on accuracy of measurement is illustrated in Tables 3 and 4. Training had no effect with criteria 2, 4, 5 and 6, but produced improved accuracy with criteria 1, 3, 7 and 8. At Observation 2, after the first training session, subjects' scores were less deviant from expert scores than at Observation 1. At Observation 3, after the second training session, many scores were more deviant than at Observation 2, yet less than Observation 1. Thus, the first training session produced greater improvement than the second training session, but in some cases the improvement was of short duration.

Criteria. Some criteria were rated much more accurately than others (Table 3). The total criterion scores indicate that criteria 3, 5, 7 and 8 ("undermined enamel removed", "axial wall depth", "pulp protection" and "margins") were rated with less deviation from expert scores than other criteria.

D. Reliability of Expert Ratings

The expert ratings were derived from the majority or mode judements of the expert raters. Between-judge reliability was calculated for the expert raters (Table 5). When total scores are considered, the experts demonstrated excellent agreement (R = .92), however, when individual criterion scores are considered, there was considerable disagreement on certain criteria. Good between-judge reliability was demonstrated in only 3 of the 8 criteria rated: "axial wall depth" (criterion 5), "pulp protection" (criterion 7), and "margins" (criterion 8). These are the same criteria which were rated accurately by the subjects.

TABLE 5

BETWEEN-JUDGE RELIABILITY OF EXPERTS'

RATINGS* (FINN TECHNIQUE)

INDIVIDUAL CRITERION SCORES								TOTAL
1	2	3	4	5	6	7	8	SCORES
.48	.47	.59	.58	.83	.45	.99	.73	.92

* Underlined values represent moderate to high reliability estimates (.70 or greater).

E. Utility

The utility of Scale C was compared with that of Scales A and B with regard to the time spent using each scale. The time subjects spent during their rating task was recorded so as to determine whether the use of a particular scale required an inordinate amount of time. After training in the use of the various scales, subjects required an average of 6-9 minutes when using Scale A, 8-9 minutes using Scale B and 11-14 minutes using Scale C. The maximum difference between these averages was 8 minutes (between 6 minutes for Scale A and 14 minutes for Scale C). This time difference was substantial, however, as these 8 minutes were spent with 40 judgements, the maximum average increased time spent using Scale C was 12 seconds per judgement.

V. DISCUSSION

Evaluation presupposes valid measures of assessment. There is little evidence, however, of the validity of current methods in dentistry. Crucial decisions concerning student promotion are based on evaluation, and it is therefore the responsibility of the rater to develop and demonstrate the validity of his measures.

There are two ways of approaching the question of reliability and validity of measurement of clinical performance in dentistry. The total scores given for students' performances or the individual components of the total scores might be analyzed. There is reason for considering both approaches. Currently in most dental schools total scores are given for students' performances and component scores are not recorded. For example, when a student works with a patient in placing a restoration he will probably achieve an A, B or C or their equivalent grade from his instructor. The student will probably not receive component grades for the many individual criteria considered by the instructor when assigning the total grade. Insofar as this method of grading clinical performance is common practice, it is necessary to consider the reliability of grading using total scores. The results of this study demonstrate that when total scores are used raters are relatively reliable and accurate when rating clinical performance.

Using total scores presupposes homogeneous items or similar criteria on which a total score is based and this is usually not the case in the measurement of clinical performance in dentistry. When the dental student is given a B grade for restoring a patient's tooth, decisions have been made concerning many different aspects of that procedure, for example, the preparation of the cavity, the use of a pulp protection, and the placement of the restorative material. The use of a high speed rotary cutting

instrument to prepare the cavity is a somewhat different operation from the use of a hand condenser to push the restorative material into the prepared cavity. From the standpoint of teaching students, it is important that students receive evaluation of individual criteria. Learning is facilitated when students receive complete and accurate feedback in regard to their performance. It is important that there be reliability of evaluation of individual criteria. The findings of this study demonstrate low reliability of evaluation of individual criteria. This low reliability could influence the efficiency of teaching clinical techniques. When a student is told by one instructor that his cavity preparation is correct and by another instructor that it is incorrect, confusion results. If instructors, and even the experts, do not agree when rating individual criteria of clinical performance, the teaching of clinical techniques cannot be efficient. Therefore, it is essential that reliability of measurement of individual criteria of clinical performance be developed, even though raters can reliably judge overall performance.

A. Extent of Clinical Experience

One of many factors which could cause low reliability in the measurement of individual criteria is the clinical background of the rater. In dentistry it is generally assumed that the raters who are best able to evaluate performance are those who are experienced performers. This is not necessarily true in other disciplines, for example, the olympic judge of skating may not necessarily be an olympic skater. Clinical experience may produce increased knowledge but it also frequently causes parochialism and rigidity, especially because dentists usually practice by themselves rather than in groups. As so much of clinical dentistry lacks operational

definition, raters tend to apply their own individual biases and do not agree when rating individual criteria of products. This was evident not only with the instructors but also with the expert raters who had little agreement when rating 5 of the 8 individual criteria. Students with little clinical experience were as reliable as instructors and no significant differences were found between students and instructors in regard to between-judge reliability and accuracy. Both students and instructors might benefit by being trained to be more reliable judges of performance. Self-evaluation which is generally not encouraged should be fostered in students early in their studies. Students who are first trained how to rate performance may better learn how to perform.

The dental assistants were generally less reliable and less accurate than the students and instructors. This finding was expected as the assistants had little if any knowledge of operative dentistry. Nevertheless, the assistants who used Scale A and Scale B benefited more from training than the students and instructors. Indeed, after training the assistants were almost as accurate in rating individual criteria as the instructors and students were before training (Table 3).

The greater benefit from training derived by the assistants may be due to the lack of preconceived biases. It might also be due to the greater amount of learning possible for the assistants; they had more to learn from training than the students or instructors. The assistants who used Scale C, however, did not benefit as much from training and this may be due to the more complex knowledge which is necessary for the use of Scale C.

B. Nature of the Rating Scale

The type of rating scale used influences agreement of raters. Although there were little differences in regard to reliability among the groups using the three scales, significant differences were evident with accuracy scores. Those who used Scale A were most accurate and most efficient indicating that a 2-point scale would be most preferable when determining competency. The use of many points in a rating scale, however, would be particularly beneficial in an instructional environment as it would provide greater feedback for the student. If many points were used in a rating scale, it would be important to define the scale points for instructional benefit. Merely increasing the number of points, without proper instruction, might serve to confuse beginning students; for example, in this study Scale C was too sophisticated for use by the assistants. Increased time would be required to use a more complicated scale, however, the time spent (20 seconds per judgement for Scale C compared with 10 seconds per judgement for Scale A) would be worthwhile considering the instructional benefits derived.

The number of points in the ideal rating scale should be a function of the number of identifiable levels of a particular characteristic. With some criteria many specified points might be necessary. With other criteria few points on the rating scale would suffice. The type of rating scale used would also depend on the use for which it would be designed. If the scale is designed for instruction, many identifiable points should be used, whereas, if it is to be used for quality control two points would be desirable. In dentistry there are many instances when the 2-point scale would be indicated. In some schools advanced senior students are placed

in honor programs whereby they practice with a minimum of instruction and instructors serve only to monitor quality. Expanded duty dental auxiliary programs are being established placing dentists in the position of "team captain" responsible for quality control. Peer review would also be a form of quality control. In these instances a 2-point scale would be most useful for both accuracy and utility.

C. Training

Simple practice with immediate-feedback training is effective in reducing measurement error and improving accuracy (see sum scores, Table 3), but it does not seem sufficient to enhance greatly within-judge or between-judge reliability. Training was most effective when administered to the assistants and this greater benefit from training may be due to the fact that, as novices, the assistants lacked preconceived biases. It might also be due to the fact that the assistants had more to benefit from training than the instructors or students. Greater improvement from training would be expected with the assistants as there was more opportunity for improvement to occur.

Training improved the reliability of evaluation of certain criteria, for some subjects, some of the time. This improvement was not always long standing. Unfortunately no trends were evident in the data to explain the shifts which occured in the magnitude of reliability indices after training.

More extensive practice with immediate-feedback training might serve to improve reliability and accuracy of rating. This type of training is dependent on the specificity of definition of scale points. When training failed it may have been because some criteria were not readily distinguishable at each of many scale points. To the extent that criteria are operationally defined, that is, defined so as to produce agreement by raters, training would be successful. The specificity which was lacking from the operational definitions contributed to some failure of training.

D. Methodological Considerations

Within-judge reliability is usually derived by taking the scores given by a judge to a group of specimens at one point in time and comparing them with scores given by the same judge to the same specimens at another point in time. In this study within-judge reliability was calculated by comparing scores given at two different points in time by many judges for eight criteria in each of two specimens. This method of calculation was used because of the small number of specimens which were graded twice by the judges. Ideally a large number of specimens should be used, and within-judge reliability should be calculated for individual judges rather than groups of judges.

Interpretations derived from overall scores should be made with caution as spurious results may occur. For example, suppose two judges were rating two criteria in one specimen and the first judge found that the first criterion was correct, whereas, the second judge found that the second criterion was correct. If individual criterion scores are considered, then no agreement exists between the two judges. However, if overall or total specimen scores are analyzed, then perfect agreement results because both judges rated the specimen with one criterion correct. The relative importance of findings based on overall scores depends on the homogeneity of criteria on which those scores are based.

E. General Implications

The fact that individual criteria appear not to be reliably measureable opens for question their relative importance as now considered in dentistry. Presently there exists mutually exclusive views in regard to certain criteria. For example, some authors suggest that internal line angles of a cavity preparation should be rounded (Finn, 1967) whereas others

suggest that these angles should be square (Howard, 1968). Clinical research must be performed to demonstrate that the angle should be round or square, or that the shape of the angle is unimportant. If the shape of the angle is important, then adequate methods to recognize the correct angle shape must be developed. Such research is essential to produce universally accepted standards of performance. Such standards are essential if student learning is to become more efficient. As appropriate measures are developed, faculty and students should be trained in their use. Faculty would then be able to provide diagnostic feedback to students in regard to performance, and students could very early in their studies practice self-evaluation. Students might more easily learn how to perform clinical procedures if they are first taught how to rate performance.

Dental practice is changing as dental auxiliaries are increasingly expected to perform expanded duties. In the future many of the technical procedures of dentistry will be delegated to auxiliaries. The dentist will have the responsibility to supervise auxiliaries and to guarantee a predetermined minimum standard of performance, as is presently done with many laboratory procedures which are now delegated to technicians. The development of reliable criterion measures should be a necessary prerequisite to the expansion of duties for auxiliary personnel. Rather than spend inordinate amounts of time learning to perform procedures which the dentist will not do, the dental student should learn to rate performance so that he is able to evaluate the work of others.

Increasingly there is discussion of developing a nationally accepted board examination of clinical performance. If such an examination is developed, it will depend on the availability of reliable and valid performance measures. In addition, if the dental profession accepts the responsibility

of peer review to continually re-examine and re-license its members, reliable and valid performance measures will be required.

This study poses questions for future research. In this study certain criteria were rated more accurately than others. No explanation for this finding is apparent. Future research could involve a more complete specification of criteria in the form of operational definitions. These might be developed through a broad based task analysis, yet they would have to be substantiated by clinical research. Research similar to this study might be performed using a more intensified amount of training. Exhaustive debriefing of raters after their rating sessions might demonstrate why raters rate as they do so that conflicts in concept could be eliminated. Training could include the use of models portraying the various points on the rating scale. With a particular standard for comparison, judges might more likely be in agreement when rating products. Future research should also include other disciplines in dentistry besides operative dentistry.

VI. SUMMARY

This study was concerned with reliability and accuracy of measurement of clinical performance in operative dentistry. The influences on reliability and accuracy of the nature of the rating scale (that is, the number and the specificity of scale points), the extent of clinical experience of the rater, and the training of raters were investigated.

Subjects included 30 instructors, 36 junior dental students, and 16 dental assistants from the University of Pittsburgh. In addition, five expert raters were used. The subject groups (instructors, students, assistants) were each subdivided into three groups so that three different rating scales could be used. The scales were: a 2-point scale with two specified

points, a 5-point scale with end points specified, and a 5-point scale with all points defined. At each of three sessions, subjects evaluated eight criteria of operative dentistry performance in five specimens, each containing one extracted mandibular second bicuspid. All scores were analyzed for between-judge and within-judge reliability, and for accuracy, that is, agreement with an expert score. Individual criterion scores as well as total test scores were used for the analysis.

The findings of this study demonstrated that at one dental school instructors were able to reliably evaluate overall performance in operative dentistry. However, instructors were not able to assess specific criteria of performance accurately. Junior dental students performed similar to instructors when rating clinical performances. Dental assistants with no experience in clinical performance of operative dentistry were able to be trained to rate overall performance fairly accurately. When scale points were defined specifically, raters tended to be more accurate in their judgements. Simple practice in the use of the rating scale with immediate feedback was effective in reducing measurement error, however, it was not sufficient to establish high reliability and accuracy of measurement.

VII. CONCLUSIONS

1. Measurement of clinical performance in operative dentistry was reliable and accurate when total performance scores were considered. It was not reliable and accurate when individual specific criteria were rated.
2. Expertise in the performance of operative dentistry was not necessary in order to produce accurate raters of clinical performance. Junior dental students were able to rate performance as reliably as instructors and dental assistants were trained to rate performance almost as accurately as instructors with no training.

3. The use of a 2-point rating scale was more accurate than the use of a 5-point rating scale. Nevertheless, a 5-point scale may be more beneficial for instructional purposes.

4. A short training session of practice with immediate feedback was not sufficient to enhance greatly the accuracy of measurement.

5. Many criteria which are thought to be of cricial importance in operative dentistry were not reliably rated by some expert raters. Clinical research is necessary to demonstrate the importance of these criteria and if they are found to be necessary, better criterion measures must be developed.

BIBLIOGRAPHY

Ebel, R.L. Estimation of the reliability of ratings. Psychometrica, 16: 407 - 424, Dec. 1951

Fernandez, J. J. Evaluation of sudent clinical performance in dental schools: construction and validation of a scale for the evaluation of cavity preparations and silver amalgam restorations in the primary dentition. PhD thesis, U. of N. Carolina, Chapel Hill, 103p. 1967.

Finn, R. H. Note on estimating the reliability of categorized data. Educational and Psychological Measurement, 20: 71 - 76, Spring, 1970.

Finn, S. B. Clinical Pedodontics, Saunders, Philadelphia, 753p. 1967.

Houpt, M. I. (1971) Accuracy of Measurement of Clinical Performance in Dentistry. PhD thesis, U. of Pittsburgh, Pittsburgh, 74p. 1971.

Howard, W. W. Atlas of Operative Dentistry, Mosby, St. Louis, 159p. 1968.

Simon, W. J. Clinical Operative Dentistry, Saunders, Philadelphia, PA. 381p. 1956.